Обучение машинного перевода без параллельных текстов*

Артеменков*[1] *А. А., Бахтеев*[1] *О. Ю., Стрижов*[2] *В. В.
[1]Московский физико-технический институт
[2]Вычислительный центр им. А. А. Дородницына ФИЦ ИУ РАН

В данной работе исследуется задача машинного перевода между двумя языками. Предлагается подход, основанный на моделях автокодировщиков и не требующий наличия большого корпуса параллельных предложений. Каждому предложению из обоих языков ставится в соответствие представление в общем скрытом пространстве. Оптимизация проводится таким образом, чтобы скрытые пространства автокодировщиков для разных языков совпадали. Для проверки качества модели проводится вычислительный эксперимент по переводу предложений между парой языков русский-украинский.

Ключевые слова: *нейронные сети, машинный перевод, автокодировщики.*

Введение

Целью данной работы является ьрешение задачи машинного перевода в отсутствии достаточного корпуса параллельных предложений. При наличии нескольких миллионов [1] параллельных образцов хорошо себя показывают методы машинного перевода с использованием нейронных сетей [2], [3]. Высокие результаты достигаются при использовании глубоких (свёрточных или рекуррентных) нейронных сетей, однако, в данном подходе критично наличие большой обучающей выборки. Частичное решение данной проблемы было найдено в пополнении числа предложений с помощью использования переводчиков более низкого качества. В [4] было показано, что данным способом могут быть улучшены результаты работы системы статистического машинного перевода Moses [5]. Более общим подходом является отказ от перевода в одну сторону и параллельное обучение переводчиков таким образом, чтобы один пополнял обучающую выборку другого.

Описанный выше метод был использована в [6] для перевода предложений с английского языка на французский. В данной работе подобная технология будет применяться для перевода с русского языка на украинский. Рассматриваются автокодировщики, реализованные в виде рекуррентных нейронных сетей [7], [8], используемые для прямого и обратного перевода, и сеть-дискриминатор, оптимизируемая с целью по представлению слова в векторном пространстве определять язык [9]. Автокодировщики оптимизируются таким образом, чтобы их латентные представления совпадали, или, что эквивалентно, чтобы дискриминатор не мог с достаточной уверенностью определить язык, соответствующий сгенерированному вектору. Для того, чтобы избежать переобучения, добавляется шум, не дающий автокодировщикам восстанавливать предложения в точности. Шаг оптимизации состоит из двух стадий: оптимизация дискриминатора и оптимизация переводчика. На первой стадии выбирается случайное предложение из исходного языка, кодируется с добавлением шума [10]) и подаётся на вход дискриминатору. После шага оптимизации аналогичные действия повторяются со случайным предложением из конечного языка. На второй стадии выбирается случайное предложение из исходного языка и переводится те-

---

Научный руководитель: Стрижов В. В. Авторы: А.В. Грабовой, О.Ю. Бахтеев, В.В. Стрижов, Eric Gaussier, координатор Малиновский Г. С. Консультант: Бахтеев О. Ю.

кущей версией переводчика на конечный язык. Затем на него накладывается шум, оно кодируется, и считываются показания дискриминатора. После шага оптимизации предложение переводится обратно в исходный язык и вычисляется значение функции потерь. После шага оптимизации действия повторяются со случайным предложением из конечного языка.

В качестве эксперимента производится перевод предложений с русского языка на украинский. Для этой пары языков отсутствует большие выборки параллельных предложений в открытом доступе, при этом достаточно данных по каждому из языков в отдельности. Качество полученного в результате переводчика оценивается с помощью метрики BLEU [11].

Постановка задачи

Введём индекс $l \in \{\text{src}, \text{tgt}\}$, показывающий, к какому языку относится рассматриваемый объект. Обозначим через $\mathfrak{D}^l = [\mathbf{s}_1^l, \ldots, \mathbf{s}_{m_l}^l]$ корпус предложений со словарём $V^l = \{w_i^l\}_{i=1}^{|V^l|}$. В общем случае, корпуса $\mathfrak{D}^{\text{src}}$ и $\mathfrak{D}^{\text{tgt}}$ не являются параллельными. Валидационная выборка, состоящая из параллельных предложений, далее будет обозначаться как $\mathfrak{D}^{valid} = \{(\mathbf{s}_1^{\text{src}}, \mathbf{s}_1^{\text{tgt}}), \ldots, (\mathbf{s}_{m_{\text{valid}}}^{\text{src}}, \mathbf{s}_{m_{\text{valid}}}^{\text{tgt}})\}$. Обозначим пространство предложений как $\mathcal{S}^l$, где каждое предложение есть набор слов: $\mathbf{s}_i^l = \{w_k\}_{k=1}^{|\mathbf{s}_i^l|}$. Введём в рассмотрение скрытое пространство $\mathcal{Z}$, общее для обоих языков, элементами которого являются коды предложений: $\{\mathbf{z}_i\}_{i=1}^k, \mathbf{z}_i \in \mathbb{R}^h$. Количество k векторов в скрытом представлении совпадает с числом слов исходного предложения и для разных предложений может отличаться.

В процессе оптимизации параметров модели каждое слово предложения представляется в виде вектора вероятностей $\mathbf{p} \in \mathbb{R}^{|V^l|}$, $\sum_{i=1}^{V^l} p_i = 1$, $p_i \geqslant 0$, где p_i есть вероятность появления слова $w_i^l \in V^l$ на данной позиции. Рассматриваемая модель состоит кодировщика $\mathbf{f}^l : \mathcal{S}^l \to \mathcal{Z}$ и декодировщика $\mathbf{g}^l : \mathcal{Z} \to \mathcal{S}^l$, отвечающих соответственно за отображение предложений из обоих языков в общее латентное пространство $\mathcal{Z}$ и обратное отображение из пространства $\mathcal{Z}$ в предложения первого или второго языка. Параметры как кодировщиков $\mathbf{f}^{\text{src}}$ и $\mathbf{f}^{\text{tgt}}$, так и декодировщиков $\mathbf{g}^{\text{src}}$ и $\mathbf{g}^{\text{tgt}}$ являются общими. Для обоих языков отличается лишь часть модели, генерирующая вектор вероятностей $\mathbf{p}$ по состоянию декодировщика.

Максимизируется следующая функция качества:

$$L = \frac{1}{m_{\text{valid}}} \sum_{i=1}^{m_{\text{valid}}} \text{BLEU}(g^{\text{tgt}}(f^{\text{src}}(\mathbf{s}_i^{\text{src}})),\ \mathbf{s}_i^{\text{tgt}})$$

Отсутствие параллельных предложений в обучающей выборке можно компенсировать с помощью сходства латентных пространств. В данной работе для этого предпринимаются следующие действия:

1. используется сеть-дискриминатор **discr**, которая по латентному представлению $\mathbf{z}$ предложения $\mathbf{s}$ определяет, какому языку оно принадлежит;
2. функция потерь записывается таким образом, чтобы оптимизировать дискриминатор **discr** для верного распознавания языка предложения, а автокодировщик – для генерации похожих представлений предложений из разных языков;
3. во избежание переобучения, каждый раз перед кодированием предложения к нему добавляется шум $\sigma(\cdot)$: из предложения опускаются некоторые слова, а к оставшимся применяется перестановка π таким образом, чтобы слова переставлялись не слишком далеко: $\max(\pi(i) - i) \leqslant k$.

Обозначим через $\mathbf{s}(k)$ вектор вероятностей $\mathbf{p}$, соответствующий k-му слову предложения $\mathbf{s}$. В качестве метрики между словами предлагается использовать $\mathrm{NLL}(\cdot,\cdot)$ (negative log likelihood).

$$\mathrm{NLL}(w_i^l, \mathbf{p}) = -\sum_{k=1}^{|V^l|} \mathbb{I}\{k=i\} \log p_k$$

Введём обозначение для противоположного языка: $\widehat{l} = \mathrm{src}$, если $l = \mathrm{tgt}$, и $\widehat{l} = \mathrm{tgt}$, если $l = \mathrm{src}$. Функции потерь для переводчика L и дискриминатора L_D записываются следующим образом:

$$L = aL_{AE} + bL_{TR} + cL_{ADV}$$

1. *Ошибка автокодировщика*
   Отвечает за правильное восстановление зашумлённых предложений из латентного пространства.
$$L_{AE} = \sum_{k=1}^{|\mathbf{s}^l|} \mathrm{NLL}(\mathbf{g}^l(\mathbf{f}^l(\sigma(\mathbf{s}^l)))(k), \mathbf{s}^l(k))$$
2. *Ошибка перевода*
   Отвечает за качество перевода с использованием автокодировщиков.
$$L_{TR} = \sum_{k=1}^{|\mathbf{s}^l|} \mathrm{NLL}(\mathbf{g}^{\widehat{l}}(\mathbf{f}^l(\sigma(\mathbf{s}^l)))(k), \mathbf{s}^l(k))$$
3. *Штраф на различие скрытых пространств*
   Отвечает за то, чтобы латентные пространства обоих автокодировщиков были похожи. Здесь же рассматривается ошибка дискриминатора L_D, показывающая, насколько точно определяется язык предложения по его скрытому представлению.
$$L_{ADV} = -\log \mathbb{P}\{\mathbf{discr}(\mathbf{f}^l(\mathbf{s}^l)) = \widehat{l}\}$$
$$L_D = -\log \mathbb{P}\{\mathbf{discr}(\mathbf{f}^l(\mathbf{s}^l)) = l\}$$

Базовый алгоритм

В качестве нулевого приближения используется модель пословного перевода. Каждое предложение пословно заменяется на ближайшего по косинусному расстоянию соседа из векторного представления MUSE [12]. Этот перевод обладает не очень высоким качеством, так как не учитывает особенности построения предложений в разных языках. Однако, после применения шума, в нём оказываются и правильно переведённые предложения. За счёт этого оптимизируемая модель имеет возможность научиться строить переводы, которые отсутствовали в нулевом приближении.

Кодировщик $\mathbf{f}^l$ является двунаправленной рекуррентной нейронной сетью GRU, декодировщик представляет собой GRU с сетью внимания и многослойную нейронную сеть, осуществляющую генерацию вектора вероятностей $\mathbf{p}$ по выходу GRU. Сеть внимания реализована с помощью однослойной нейронной сети, а дискриминатор – многослойной нейронной сети. Перед подачей на вход кодировщику все слова в предложении заменяются на их векторное представление, которое инициализируется векторами MUSE и оптимизируется вместе с другими параметрами.

Алгоритм оптимизации может быть записан следующим образом:

1. Построить нулевое приближение переводчика M
2. Пока не достигнуто желаемое качество перевода:
   (а) Сделать T шагов соревновательной оптимизации:
      i. Вычислить L и L_D
      ii. Обновить параметры автокодировщиков с помощью градиентного спуска для L при фиксированных параметрах дискриминатора
      iii. Обновить параметры дискриминатора с помощью градиентного спуска для L_D при фиксированных параметрах автокодировщиков
   (б) Обновить M, записав в него переводчик, полученный из последовательного применения кодировщика и декодировщика
   (в) Построить новый перевод обучающей выборки с использованием M

Вычислительный эксперимент

В качестве обучающей выборки использовались данные Multi30k [13], [14]. Параметры автокодировщиков оптимизировались с помощью Adam, параметры дискриминатора – с помощью RMSprop.

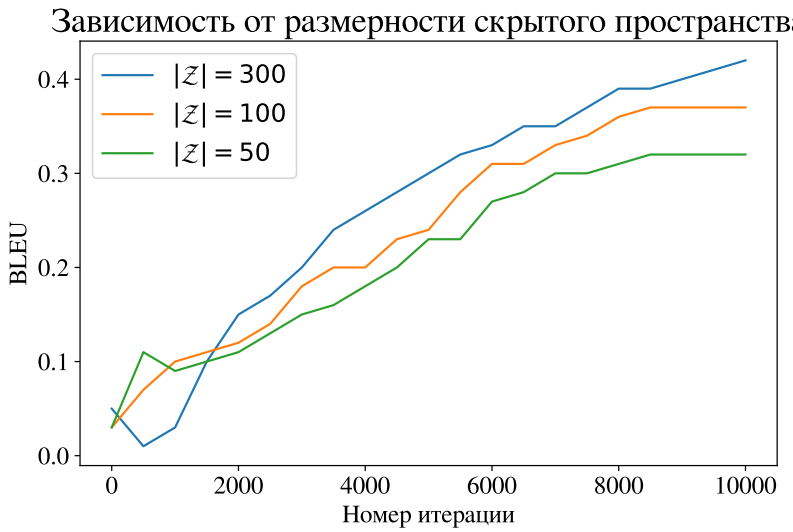

Рис. 1. Зависимость BLEU от номера итерации в зависимости от размерности скрытного пространства.

Вывод

В данной работе была исследована задача машинного перевода между двумя языками. Предложен подход, основанный на моделях автокодировщиков и не требующий наличия большого корпуса параллельных предложений. Работа была проделана на англо-французской выборке, так как качество найденных русско-украинских выборок не является достаточным как для построения хорошей модели, так и для валидации: в предложениях присутствует слишком много слов не из словаря (числа и формы слов), а сами переводы не всегда точны. Возможна дальнейшая работа по модификации алгоритма таким образом, чтобы это не так сильно влияло на качество перевода [15], [16]. В данных работах предлагается извлечь дополнительную информацию из выборки для одного языка, что позволит сохранить независимость метода от наличия параллельных предложений.

Литература

[1] Bahdanau Dzmitry, Cho Kyunghyun, Bengio Yoshua. Neural machine translation by jointly learning to align and translate // arXiv preprint arXiv:1409.0473. 2014.

[2] On the properties of neural machine translation: Encoder-decoder approaches / Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau [и др.] // arXiv preprint arXiv:1409.1259. 2014.

[3] Luong Minh-Thang, Pham Hieu, Manning Christopher D. Effective approaches to attention-based neural machine translation // arXiv preprint arXiv:1508.04025. 2015.

[4] Bertoldi Nicola, Federico Marcello. Domain adaptation for statistical machine translation with monolingual resources // Proceedings of the fourth workshop on statistical machine translation / Association for Computational Linguistics. 2009. С. 182–189.

[5] Moses: Open source toolkit for statistical machine translation / Philipp Koehn, Hieu Hoang, Alexandra Birch [и др.] // Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions / Association for Computational Linguistics. 2007. С. 177–180.

[6] Lample Guillaume, Denoyer Ludovic, Ranzato Marc'Aurelio. Unsupervised machine translation using monolingual corpora only // arXiv preprint arXiv:1711.00043. 2017.

[7] Gers Felix A, Schmidhuber Jürgen, Cummins Fred. Learning to forget: Continual prediction with LSTM. 1999.

[8] Graves Alex, Schmidhuber Jurgen. Framewise phoneme classification with bidirectional LSTM and other neural network architectures // Neural Networks. 2005. Т. 18, № 5-6. С. 602–610.

[9] Goldberg Yoav, Levy Omer. word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method // arXiv preprint arXiv:1402.3722. 2014.

[10] Kim Yunsu, Ney Jiahui Geng Hermann. Improving Unsupervised Word-by-Word Translation with Language Model and Denoising Autoencoder. 2018.

[11] BLEU: a method for automatic evaluation of machine translation / Kishore Papineni, Salim Roukos, Todd Ward [и др.] // Proceedings of the 40th annual meeting on association for computational linguistics / Association for Computational Linguistics. 2002. С. 311–318.

[12] Word translation without parallel data / Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato [и др.] // arXiv preprint arXiv:1710.04087. 2017.

[13] Findings of the Second Shared Task on Multimodal Machine Translation and Multilingual Image Description / Desmond Elliott, Stella Frank, Loïc Barrault [и др.] // Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers. Copenhagen, Denmark: Association for Computational Linguistics, 2017. September. С. 215–233. URL: http://www.aclweb.org/anthology/W17-4718.

[14] Multi30K: Multilingual English-German Image Descriptions / Desmond Elliott, Stella Frank, Khalil Sima'an [и др.] // Proceedings of the 5th Workshop on Vision and Language. Association for Computational Linguistics, 2016. С. 70–74. URL: http://www.aclweb.org/anthology/W16-3210.

[15] Irvine Ann, Callison-Burch Chris. End-to-end statistical machine translation with zero or small parallel texts // Natural Language Engineering. 2016. Т. 22, № 4. С. 517–548.

[16] Toward statistical machine translation without parallel corpora / Alexandre Klementiev, Ann Irvine, Chris Callison-Burch [и др.] // Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics / Association for Computational Linguistics. 2012. С. 130–140.